MODELS AND METHODS FOR CLUSTERWISE LINEAR REGRESSION

C. Hennig

Institut für Mathematische Stochastik,
Universität Hamburg, Bundesstr. 55, D-20146 Hamburg, Germany

Abstract: Three models for linear regression clustering are given, and corresponding methods for classification and parameter estimation are developed and discussed: The mixture model with fixed regressors (ML-estimation), the fixed partition model with fixed regressors (ML-estimation), and the mixture model with random regressors (Fixed Point Clustering). The number of clusters is treated as unknown. The approaches are compared via an application to Fisher's Iris data. By the way, a broadly ignored feature of these data is discovered.

1 Introduction

Cluster analysis problems based on stochastic models can be divided into two classes:

1. A cluster is considered as a subset of the data points, which can be modeled adequately by a distribution from a class of *cluster reference distributions* (c.r.d.). These distributions are chosen to reflect the meaning of *homogeneity* with respect to the certain data analysis problem. Therefore c.r.d. are often unimodal. If the class of c.r.d. is parametric, then one is interested in classification of the data points and parameter estimation within each cluster.

2. A cluster is considered as an area of high density of the distribution of the whole dataset. No distributional assumption is made for the single clusters.

Clusterwise linear regression is a problem of the first kind since the points of each cluster are considered to be generated according to some linear regression relation, i.e. one imagines a separate model for each cluster. The class of c.r.d. for the regression clustering problem contains distributions of the following kind: Consider a dataset $\mathbf{Z} = (x_i', y_i)_{i \in I}$, $x_i \in \{1\} \times \boldsymbol{R}^p, y_i \in \boldsymbol{R}$, I being some index set.

$$\begin{gathered}\mathcal{L}(y_i|x_i) = F_{(x_i,\beta,\sigma^2)}, \text{ defined by} \\ y_i = x_i'\beta + u_i, \qquad \mathcal{L}(u_i) = \mathcal{N}_{(0,\sigma^2)}, \\ (\beta, \sigma^2) \in \boldsymbol{R}^{p+1} \times \boldsymbol{R}_0^+.\end{gathered}$$

The first component of β denotes the intercept. The $u_i, i \in I$ are considered to be stochastically independent. The x_i are called *regressors* in the following. They can be fixed or random with $\mathcal{L}(x_i) = G$ from some class of distributions $\mathcal{G}$. In the latter case the regressors are assumed to be i.i.d. and independent of $(u_i)_{i\in I}$. F_{G,β,σ^2} then denotes the joint distribution of (x_i, y_i). In our setup, all parameters are considered as unknown.

The models will be divided into fixed and random regressor models, and into mixture and fixed partition models. Mixture models treat the cluster membership of a point as random, fixed partition models contain parameters for the cluster membership of each point. A fixed partition model with random regressors will not be given because this does not lead to an easy clustering method. The purpose of the model based approach presented here is not to describe the mechanism generating the data, but to find an adequate description of the data themselves. Thus, all models can be applied to the same data. In particular, the question is ignored if the regressors were *really* fixed or random.

The literature on clusterwise linear regression either treats the mixture model with fixed regressors (e.g. Quandt and Ramsey (1978), for general p and number of clusters DeSarbo and Cron (1988)) or discusses algorithms for a least squares solution (e.g. Bock (1969), Spaeth (1979)) which is related to the fixed regressors fixed partition model presented here in the case of equal error variances for each cluster. This paper is based on the unpublished dissertation Hennig (1997b) where simulation results and proofs are given in full detail.

2 Fixed regressors, mixture model

Let I be an index set, usually $I = \{1, \ldots, n\}$. With a given regressor design $(x_i)_{i\in I} \in (\{1\} \times \mathbb{R}^p)^I$ the **fixed regressors mixture model** (FRM) is defined by

$$\mathcal{L}((y_i)_{i\in I}) = \bigotimes_{i\in I} \sum_{j=1}^{s} \epsilon_j F_{(x_i,\beta_j,\sigma_j)},$$

$$\sum_{j=1}^{s} \epsilon_j = 1,\ \epsilon_j > 0,\ j = 1, \ldots, s.$$

That is, s denotes the number of clusters and ϵ_j denotes the proportion of the cluster j. The log-likelihood function

$$\ln L_n(s, (\epsilon_j, \beta_j, \sigma_j^2)_{j=1,\ldots,s}, \mathbf{Z}) =$$
$$\sum_{i\in I} \ln\left(\sum_{j=1}^{s} \epsilon_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[\frac{-(y_i - \beta_j' x_i)^2}{2\sigma_j^2}\right]\right)$$

can be locally maximized for given s via the EM-algorithm described in DeSarbo and Cron (1988). This works only subject to $\sigma_j^2 > c\ \forall j$ with some lower bound $c > 0$ (e.g. $c = 0.001$) since otherwise $\ln L_n$ would be unbounded. After having performed the algorithm, point i can be classified to cluster $\hat{\gamma}(i) \in \{1, \ldots, s\}$ according to

$$\hat{\gamma}(i) = \arg\max_j \hat{\epsilon}_{ij}, \quad \hat{\epsilon}_{ij} = \frac{\hat{\epsilon}_j \varphi_{(x_i'\beta_j, \hat{\sigma}_j^2)}(y_i)}{\sum_{l=1}^s \hat{\epsilon}_l \varphi_{(x_i'\beta_l, \hat{\sigma}_l^2)}(y_i)}.$$

$\hat{\epsilon}_{ij}$ denotes the estimated a posteriori probability for point i to be generated by mixture component j.

The consistency proofs for FRM-ML estimation (Kiefer (1978), DeSarbo and Cron (1988)) suffer from not taking possible identifiability problems (Hennig (1996)) into account.

DeSarbo and Cron (1988) suggest Akaike's Information Criterion (AIC) for the estimation of s:

$$\hat{s} := \arg\max_s \ln \hat{L}_n(s) - k(s), \; k(s) = (p+3)s - 1.$$

$k(s)$ denotes the number of free parameters to estimate for the cluster number s and $\ln \hat{L}_n(s)$ is the estimated maximum log-likelihood. Their simulations do not treat the performance of this proposal. The simulations of Hennig (1997b) show the tendency of the AIC to overestimate a small number of clusters. Schwarz' Criterion (SC) gives smaller estimates of s for $n > e^2$ and seems to work better:

$$\hat{s} := \arg\max_s \ln \hat{L}_n(s) - \frac{\ln n}{2} k(s).$$

The discussion of the Iris data example in section 5 illustrates this performance. Up to now there are no theoretical results on the performance of the AIC and SC for linear regression mixtures.

Some alternative proposals for parameter estimation within this model were made (e.g. Quandt and Ramsey (1978)), but they lead to greater numerical difficulties and were investigated only for $s = 2, p = 1$.

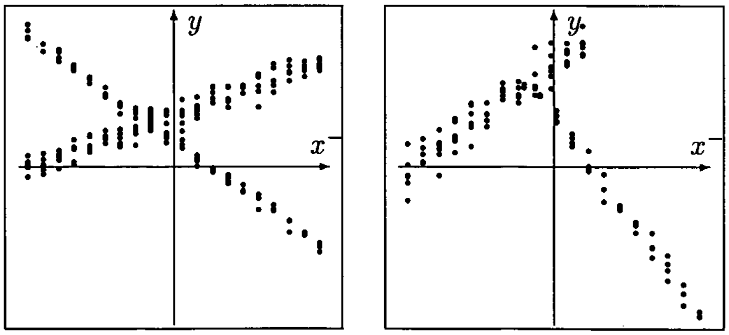

Figure 1: Assignment independence - assignment dependence

The implicit assumption of **assignment independence** is a disadvantage of the FRM. That is, the clusters keep the same proportions $\epsilon_j, j = 1, \ldots, s$ for every fixed regressor x_i (see figure 1). The probability of a point (x_i, y_i) to be generated by cluster j is independent of x and i. This is not generally true. For example in a **change point** setup, the cluster membership is considered as *determined* by x or i. Methods concerning this particular assumption can be found e.g. in Krishnaiah and Miao (1988). Also for the Iris data in section 5, assignment independence seems not to be fulfilled.

3 Fixed regressors, fixed partition model

In the fixed partition approach, the cluster membership of each point i is indicated by a parameter $\gamma(i)$. Thus, general kinds of assignment dependency can be modeled. The **fixed regressors fixed partition model** (FRFP) is given by

$$\begin{aligned} \mathcal{L}((y_i)_{i\in I}) &= \bigotimes_{i\in I} F_{\left(x_i,\beta_{\gamma(i)},\sigma^2_{\gamma(i)}\right)}, \\ \gamma:\ I &\mapsto \{1,\ldots,s\}, \end{aligned}$$

$(x_i)_{i\in I} \in (\boldsymbol{R}^{p+1})^I$ again given fixed. Under known s, ML-estimation is also possible within this model. The log-likelihood function is given by

$$\begin{aligned} &\ln L_n(s,\gamma,(\beta_j,\sigma_j^2)_{j=1,\ldots,s},\mathbf{Z}) = \\ &-\tfrac{1}{2}\sum_{j=1}^{s}\sum_{\gamma(i)=j}\left(\ln 2\pi + \ln\sigma_j^2 + \frac{(y_i-\beta_j' x_i)^2}{\sigma_j^2}\right). \end{aligned} \qquad (1)$$

For given $(\hat\beta_j, \hat\sigma_j^2)_{j=1,\ldots,s}$, (1) is maximized according to

$$\hat\gamma(i) = \arg\min_j \left(\ln\hat\sigma_j^2 + \frac{(y_i-\hat\beta_j' x_i)^2}{\hat\sigma_j^2}\right). \qquad (2)$$

For given $\hat\gamma$, (1) is the sum of the usual log-likelihood functions for homogenous linear regressions within each cluster. Therefore, it is minimized by the LS-estimator $\hat\beta_j$ from the points (x_i, y_i) with $\hat\gamma(i) = j$ and

$$\hat\sigma_j^2 := \frac{\sum_{\hat\gamma(i)=j}(y_i-\hat\beta_j' x_i)^2}{\hat n_j}, \quad \hat n_j := \sum_{i=1}^{n} 1(\hat\gamma(i)=j),\ j=1,\ldots,s. \qquad (3)$$

That is, $\ln\hat L_n$ is monotonely increased if the steps (2) and (3) are carried out alternately. This algorithm leads to a local maximum in finitely many steps since there are only finitely many choices for $\hat\gamma$. In my experience, this is noticeably the fastest algorithm discussed in this paper. Under $\sigma_1^2 = \ldots = \sigma_s^2$, the procedure is equivalent to the least squares algorithm of Bock (1969).

There is some literature that compares mixture and fixed partition approaches applied to location-scale and especially Gaussian distributed clusters (e.g. Bryant and Williamson (1986)). Analogously to the location-scale case it can be shown that FRFP-ML leads to inconsistent parameter estimators. This does not matter in practice if the clusters are well separated, but causes serious problems otherwise. Like FRM-ML, FRFP-ML needs some lower bound on the error variance parameters since otherwise $\ln \hat{L}_n$ would be unbounded.

The approaches for the estimation of s discussed in section 2 are not reasonable here because the number of parameters $\gamma(i)$ increases with n and their value range increases with s. The following modification of the SC worked very well in simulations:

$$\hat{s} := \arg\max_s \ln \hat{L}_n(s) - \frac{\ln n}{2} k(s) - 0.7sn, \quad k(s) := (p+2)s, \tag{4}$$

$k(s)$ denoting the number of regression and scale parameters.

4 Random regressors, mixture model

Random regressors have the advantage that the observations can be treated as i.i.d. The **random regressors mixture model** (RRM) has the following form: $(x_i, y_i) \in \{1\} \times \boldsymbol{R}^p \times \boldsymbol{R},\ i \in I$, are distributed i.i.d. according to

$$\mathcal{L}(x,y) = \sum_{j=1}^{s} \epsilon_j F_{(G_j,\beta_j,\sigma_j^2)},$$
$$\sum_{j=1}^{s} \epsilon_j = 1,\ G_1, \ldots, G_s \in \mathcal{G},$$

that is, $\mathcal{L}(x) = G_j$ within cluster j. Suitable choices for $G_j,\ j = 1, \ldots, s$, enable us to model every kind of assignment (in-)dependence. Usually the G_j are not of interest, but unknown. For performing ML-estimation, there needs to be a parametric specification of $\mathcal{G}$. This will not be discussed here. A more general approach is presented instead. The RRM is a special case of the **contamination model** (CM) (choose $F^* = \sum_{j=2}^{s} \frac{\epsilon_j}{\epsilon} F_{(G_j,\beta_j,\sigma_j^2)}$, $(G, \beta, \sigma^2) = (G_1, \beta_1, \sigma_1^2)$, $\epsilon = \epsilon_1$ below):

$$\mathcal{L}(x,y) = (1-\epsilon) F_{(G,\beta,\sigma^2)} + \epsilon F^*,\ 0 \leq \epsilon < 1,\ G \in \mathcal{G}. \tag{5}$$

There is some basic difference between the CM and the former models. The parameters (G, β, σ^2) are clearly not unique in (5) since they can correspond to $(G_j, \beta_j, \sigma_j^2)$ of the RRM for each j. Further, if F^* is not assumed to be of a mixture type, the CM allows for outliers, i.e. points in the data, which do not belong to any regression population. In robust statistics, the CM with $\epsilon < \frac{1}{2}$ is a standard tool to describe the occurence of outliers.

A method to analyze the CM should find possible choices for (β, σ^2) (G is treated as nuisance) and therefore needs no specification of some number of clusters.

This goal can be achieved by means of **Fixed Point Clustering**. The idea of this approach is that a data subset, which contains no outliers, can be viewed as homogeneous. If at the same time all other points of the dataset are outliers with respect to the subset, then the subset is separated from the rest and can be considered as a cluster.

For an indicator vector $g \in \{0,1\}^n$ define $\mathbf{Z}(g) := (x_i', y_i)_{g_i=1}$.

Definition: $\mathbf{Z}(g)$ is called *Fixed Point Cluster* (FPC) w.r.t. $\mathbf{Z}$, iff g is a fixed point of

$$f: \{0,1\}^n \mapsto \{0,1\}^n,$$
$$f_i(g) = 1\left[(y_i - x_i'\hat{\beta}(\mathbf{Z}(g)))^2 \leq c\hat{\sigma}^2(\mathbf{Z}(g))\right]$$

with some prechosen constant c (e.g. $c = 10$). $\hat{\beta}(\mathbf{Z}(g))$ and $\hat{\sigma}^2(\mathbf{Z}(g))$ are regression parameter and error variance estimators based only on the data subset $\mathbf{Z}(g)$, e.g. the ML-estimators from (3).

The function f is an inversed outlier identifier (0 for outliers) based on the random regressor linear regression model. That is, a point is considered as an outlier w.r.t. $F_{(G,\beta,\sigma^2)}$ if it falls into the outlier region $\{(y - x'\beta)^2 > c\sigma^2\}$ (see Davies and Gather (1993) for the concept of model based outlier regions). Therefore an FPC $\mathbf{Z}(g)$ is exactly the set of non-outliers in $\mathbf{Z}$ w.r.t. $\mathbf{Z}(g)$ and can be interpreted as the set of "ordinary observations" generated by some member of the c.r.d.-family.

The method is similar to some procedures for robust regression where the goal is to find a solution of $\sum \rho(\frac{(y_i - x_i'\beta)^2}{\sigma^2}) \overset{!}{=} \min_\beta$. The function ρ also provides some kind of outlier identification. Local minima could be interpreted as parameters for clusters (Morgenthaler (1990)), but the choice of σ^2 is not clear and a robust estimator would depend on at least half of the data. This is not adequate for cluster analysis.

FPCs can be computed with the usual fixed point algorithm ($g^{j+1} = f(g^j)$) which converges in finitely many steps (proven in Hennig (1997b)). In order to find all relevant clusters, the algorithm must be started many times with various starting vectors g. A complete search is numerically impossible. However, this also holds for the other two methods unless one is satisfied with a local maximum of unknown quality of the log-likelihood function.

The FPC methodology does not force a partition of the dataset. Non-disjoint FPCs and points are possible, which do not belong to any FPC. According to that, FPCs are rather an exploratory tool than a parameter estimation procedure in the case of a valid partition or mixture model.

The application of FPC analysis to more general situations is discussed in Hennig (1997a), Hennig (1998).

5 Iris data example and comparison

Fisher's Iris data (Fisher (1936)) consists of four measurements of three species of Iris plants. The measurements are sepal width (SW), sepal length (SL), petal width (PW) and petal length (PL). The species are Iris setosa (empty circles in figure 2a), Iris virginica (filled circles) and Iris versicolor (empty squares). Each species is represented by 50 points. Originally, the classification of the plants was no regression problem. The dataset is used for illustratory purposes here. Find a more "real world" but less illustratory example in Hennig (1998). Only the variables SW and PW are considered. PW is modeled as dependent of SW. The distinction in "regressor" and "dependent variable" is artificial. The methods use no information about the real partition. By eye, the setosa plants are clearly seperated from the other two species, while virginica and versicolor overlap. A linear regression relation between SW and PW seems to be appropriate within each of the species.

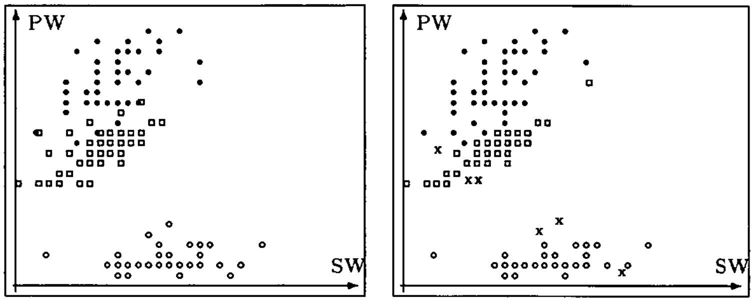

Figure 2: Iris data: a) original species - b) FRM-ML clusters with SC

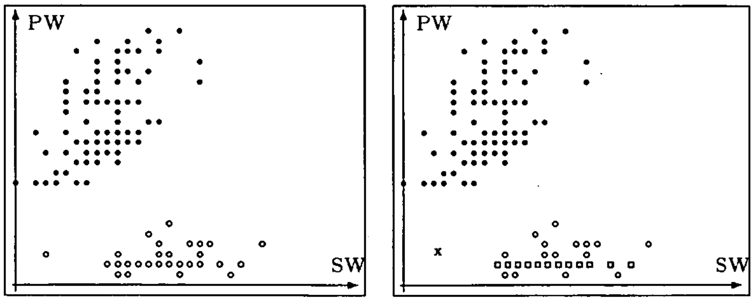

Figure 3: a) FRFP-ML clusters - b) Fixed Point Clusters

Using the SC for estimating the number of clusters, FRM-ML-estimation finds the four clusters shown in figure 2b. Three clusters correspond to the three species. FRM-ML is the only method which provides a rough distinction between the virginica and versicolor plants. The fourth cluster (crosses in figure 2b) is some kind of "garbage cluster". It contains some points which are not fitted good enough by one of the other three regression equations. Note that the deviation from assignment independence of the four cluster solution seems to be lower than that of the original partition of the species. The AIC for estimating the number of cluster leads to five clusters by removing further points from the three large clusters and building a second garbage cluster.

By application of (4), the number of clusters is estimated as 2. Figure 3a shows the ML-classification using the FRFP. It corresponds to the most natural[1] eye-fit. The well separated setosa plants form a cluster, the other two species are put together.

With 150 randomly chosen starting vectors, four FPCs are found. The first contains the whole dataset. This happens usually and is an artifact of the method. One has to know that to interpret the results adequately. The second and third cluster correspond to the setosa plants and the rest of the data, respectively. The point labelled by a cross falls in the intersection of both clusters and is therefore indicated as special. The fourth cluster is labelled by empty squares and consists of 29 points from the setosa cluster, which lie exactly on a line because of the rounding of the data. The other methods are not able to find this constellation because of the lower bounds on the error variances.

After having noticed this result, one realizes that there are other groups of points, which lie exactly on a line, and which are not found by the random search of Fixed Point Clustering since they are too small. The fourth FPC contains more than half of the setosa species[2] and is therefore a remarkable feature of the Iris data.

The results from the Iris data highlight the special characteristics of the three methods. The simulation study of Hennig (1997b) leads to similar conclusions.

FRM-ML-estimation is the best procedure if assignment independence holds and if the clusters are not well separated. At the Iris data, it can discriminate between virginica and versicolor. The stress is on regression and error variance parameter estimation.

FRFP-ML-estimation is the best procedure under most kinds of assignment dependence to find well separated clusters if there is a clear partition of the dataset. At the Iris data, the procedure finds the visually clearest constellation. The stress is on point classification.

Fixed Point Clustering is the best procedure to find well separated clusters if outliers or identifiability problems (Hennig (1996)) exist. Its stress is on exploratory purposes. By means of Fixed Point Clustering, the discovery that a large part of the setosa cluster lies exactly on a line was made.

[1] It is not clear, what "most natural" means, but this is the impression of the author.

[2] One cannot see 29 squares because some of the points are identical.

References

BOCK, H.H. (1969): The equivalence of two extremal problems and its application to the iterative classification of multivariate data. Lecture note, Mathematisches Forschungsinstitut Oberwolfach.

BRYANT, P. G., and WILLIAMSON, J. A. (1986): Maximum likelihood and classification: A comparison of three approaches. In: Gaul, W., and Schader, W. (Eds.): *Classification as a Tool of Research*, Elsevier, Amsterdam, 35-45.

DAVIES, P. L., and GATHER, U. (1993): The identification of multiple outliers. *Journal of the American Statistical Association 88, 782-801.*

DESARBO, W. S., and CRON, W. L. (1988): A Maximum Likelihood Methodology for Clusterwise Linear Regression. *Journal of Classification 5, 249-282.*

FISHER, R. A. (1936): The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7, 179-184.*

HENNIG, C. (1996): Identifiability of Finite Linear Regression Mixtures. Preprint No. 96-6, Institut für Mathematsche Stochastik, Universität Hamburg.

HENNIG, C. (1997a): Fixed Point Clusters and their Relation to Stochastic Models. In: Klar, R., and Opitz, O. (Eds.): *Classification and Knowledge Organization*, Springer, Berlin, 20-28.

HENNIG, C. (1997b): Datenanalyse mit Modellen für Cluster linearer Regression. Dissertation, Institut für Mathematsche Stochastik, Universität Hamburg.

HENNIG, C. (1998): Clustering and Outlier Identification: Fixed Point Cluster Analysis. In: Rizzi, A., Vichi, M., and Bock, H.-H. (Eds.): *Advances in Data Science and Classification*, Springer, Berlin, 37-42.

KIEFER, N. M. (1978): Discrete parameter variation: Efficient estimation of a switching regression model. In: *Econometrica 46, 427-434.*

KRISHNAIAH, P. R., MIAO, B. Q. (1988): Review about estimation of change points. In: Krishnaiah, P. R., and Rao, P. C. (Eds.): *Handbook of Statistics, Vol. 7*, Elsevier, Amsterdam, 375-402.

QUANDT, R. E., RAMSEY, J. B. (1978): Estimating mixtures of Normal distributions and switching regressions. *Journal of the American Statistical Association 73, 730-752.*

SPAETH, H. (1979): Clusterwise linear regression. *Computing 22, 367-373.*